Analysis of phone errors attributable to phonological effects associated with language acquisition through bottleneck feature visualisations

Fringi, Evangelia; Russell, Martin

DOI:
10.21437/Interspeech.2018-2422

Interspeech 2018
2-6 September 2018, Hyderabad

Analysis of phone errors attributable to phonological effects associated with language acquisition through bottleneck feature visualisations

Eva Fringi[1], *Martin Russell*[1]

[1]Department of Electronic Electrical and Systems Engineering,
University of Birmingham B15 2TT, UK

exf111@bham.ac.uk, M.J.RUSSELL@bham.ac.uk

Abstract

Previous work aimed to investigate the extent to which errors attributable to phonological effects associated with language acquisition (PEALA) contribute to the output of children's ASR. Opposite to what was intuitively expected, the proportion of errors predictable from PEALA was positively correlated with recognition accuracy, therefore increased across ages. In order to interpret this finding, the present paper employs a DNN-HMM automatic speech recognition system, built on the CSLU children's speech corpus, to produce bottleneck feature (BNF) visualisations of phones and examine how these relate with respect to PEALA. The focus is drawn particularly on ASR errors caused by phone confusions, which are compared against phone substitution pairs indicated by PEALA. The ASR results confirm the previously observed interaction between errors predictable from PEALA and rising accuracy, but also suggest that these errors only account for a small percentage of the total phone substitution error. The BNF visualisations for the most part outline the age progression smoothly and demonstrate clear clusters of neighbouring phones consistently. The distance between PEALA related phones can be partitioned in four sets; two that increase with age (at a higher or lower rate), one that roughly remains constant and one that decreases with age.

Index Terms: speech recognition,children's speech, bottleneck features

1. Introduction

Children's speech is characterised by high variability both in terms of acoustic [1], [2], [3] and linguistic components [4]. As a result, by comparison with adults', children's automatic speech recognition (ASR) yields low accuracy rates. In previous work [5] it was attempted to explore the role of linguistic variability in the performance of children's ASR by focusing on phonological effects associated with language acquisition (PEALA). This paper builds up on the previous findings and offers a visual representation of PEALA through the use of bottleneck features (BNF).

As indicated by studies on speech development [6], [7], children demonstrate systematic error patterns in early speech productions which are termed 'phonological processes'. They involve alterations of the adult target form through reductive, assimilatory or substituting mispronunciations. These are part of language acquisition and are expected to have faded out by the age of six [8], [9], [10]. However, there is evidence that both perceptually [11], [12] and physically [13], [14], [15] speech mechanisms continue their maturing processes well into adolescence.

On these grounds it has been hypothesized that beyond the age of six some vestige of PEALA, even though indiscriminable to a human listener, might still be detected by ASR systems and cause phone confusions. Following a literature review of normative studies on speech development of native english speakers [8], [9], [10], [16], [17], a set of possible ASR phone substitutions that could be predicted from PEALA has been introduced in Table 1 [18].

Table 1: *Predictable Substitutions based on PEALA.*

Voicing	Stopping	Fronting
$/p/ \rightarrow /b/$ $/t/ \rightarrow /d/$ $/k/ \rightarrow /g/$ $/s/ \rightarrow /z/$	$/s/ \rightarrow /t/, /v/ \rightarrow /b/$ $/f/ \rightarrow /p/, /th/ \rightarrow /p/$ $/jh/ \rightarrow /d/, /v/ \rightarrow /p/$ $/ch/ \rightarrow /t/, /dh/ \rightarrow /d/$ $/sh/ \rightarrow /t/, /s/ \rightarrow /th/$	$/k/ \rightarrow /t/$ $/g/ \rightarrow /d/$ $/g/ \rightarrow /t/$ $/sh/ \rightarrow /s/$
Deaffrication	**Fricative Simplification**	**Gliding**
$/ch/ \rightarrow /sh/$ $/jh/ \rightarrow /zh/$ $/ch/ \rightarrow /k/$ $/zh/ \rightarrow /z/$	$/th/ \rightarrow /f/$	$/r/ \rightarrow /w/$ $/r/ \rightarrow /l/$ $/l/ \rightarrow /w/$ $/l/ \rightarrow /y/$

In earlier work, two relatively small speech corpora comprising recordings of American English spoken by children aged between five and nine years were utilised to build a couple of baseline GMM-HMM systems and perform phone recognition. In both cases, the results suggested a positive correlation between the proportion of phone confusions predictable from PEALA and recognition accuracy. Thus, in a counter-intuitive manner, as speaker ages increased, errors attributable to PEALA became more prominent [5].

The present paper aims to interpret this observation and provide a closer understanding of its underlying causes. On this end, a larger children's speech corpus ranging across a broader age span is employed, to train both GMM-HMM and DNN-HMM based systems and offer an extensive analysis of phone confusions predictable from PEALA.

Moreover, an attempt is made to examine the relationships within PEALA related phone pairs in the acoustic domain and how these evolve with speaker age progression. This is facilitated by the extraction of multidimensional bottleneck features of the training set, which are then plotted in two-dimensional graphs after linear discriminant analysis (LDA) processing. The resulting images reflect the similarities in the acoustic models of several phones through their proximity in the BNF plane.

The next section contains a description of the methods used in the study. Section 3 provides a summary of the obtained ASR results and section 4 presents the BNF visualisation findings. Finally, section 5 contains the conclusions of the paper.

10.21437/Interspeech.2018-2422

2. Method

2.1. Data Set

The data used in this study is a subset of the CSLU corpus [19] containing American English recordings from 1116 students from Portland, Oregon. The speaker range spans from kindergarteners (5 year-olds) to tenth graders (15 year-olds) randomly allocated either to the train (665 speakers) or test set (451 speakers). The recordings feature single word or short sentence repetitions which amount to approximately 25 hours for the training set and 17 hours for the test set. A total of 47,532 utterances were automatically transcribed at the phone level according to the 39 phone set of the CMU pronunciation dictionary, using forced alignment applied to the word level transcriptions that are provided with the data.

2.2. ASR Systems

All ASR systems described below were developed with the use of The Kaldi Speech Recognition Toolkit [20] and were trained using age independent data. A 'flat' phone-loop grammar was applied in each case.

2.2.1. GMM-HMM

GMM1: This is the first, baseline recogniser. The speech was transformed into sequences of 39 dimensional feature vectors comprising 12 mel frequency cepstral coefficients (MFCCs) plus C0, augmented with the corresponding Δ and Δ^2 parameters. The system uses 1951 physical states each associated with a Gaussian mixture model (GMM) whose components are chosen from a set of 15,050 shared Gaussian PDFs.

GMM2: This recogniser was developed from 'GMM1' after applying maximum likelihood linear transform (MLLT), linear discriminant analysis (LDA) and speaker adaptive training (SAT) to obtain 40 dimensional feature vectors. The resulting system has 1997 physical states each associated with a Gaussian mixture model (GMM) whose components are chosen from a set of 15,022 shared Gaussian PDFs. Finally, forced alignment is applied to obtain an alignment between the data and the 1997 senones (GMM-HMM states). This alignment is passed to the next stage.

2.2.2. DNN-HMM

DNN1: The initial DNN-HMM system was built using the previously trained 40 dimensional fMLLR features, which were created in the SAT stage of 'GMM2' development, and the alignment from the GMM2 system. The inputs to the DNN are feature vectors in context, with a context of ± 5 frames. The number of hidden layers used was 2 and the hidden layer dimensions were 1024. Thus the DNN can be characterized as $440 \times 1024 \times 1024 \times 1997$. DNN parameter estimation used 6 iterations of state-level minimum Bayes risk (sMBR) training.

DNN2: Alignments from 'DNN1' were used to train a new DNN which includes a 9 dimensional bottleneck layer in addition to the two existing 1024 dimensioan hidden layers. The extracted BNFs were used instead of fMLLR features to train another DNN recogniser, following the same procedure as for 'DNN1'. The choice of 9 dimensions for the BNFs was based on the results presented in [21], which show that the performance using 9 BNFs is comparable with the phone recognition accuracy obtained with standard 39 dimensional MFCC-based feature vectors.

In addition to their use as feature vectors for speech recognition, BNFs are of interest because of their utility for visualization of speech signals [22]. This is investigated in Section 4.

2.3. Process

Phone recognition accuracy results were obtained from each ASR system, as well as phone confusion matrices. The set of 27 phone substitution pairs presented in Table 1 was used as a reference for determining which substitutions were predictable from PEALAs.

A significant issue in this work is the fact that the phone-level transcriptions are obtained from forced alignment using baseform transcriptions from the CMU dictionary. If a child makes systematic pronunciation errors then these will occur in both the training and test sets.

To address this issue, the assumption was made that if there are children in the training set who exhibit a particular phonological effect, then the models for the corresponding phones will be corrupted. For example if a child uses $/t/$ for $/k/$, the $/k/$ phone models will tend to be more $/t/$-like and so there will be an increase of $/t/ \rightarrow /k/$ substitutions in the test. To cater for that implication, we looked at both directions of confusion for each of the 27 effects: the proportion of each substitution error predictable from PEALA was calculated as the ratio of confusions in both directions of each pair, over the total confusions for the phones involved:

$$\text{Proportion of } p_1 \rightarrow p_2 = \frac{p_1 \rightarrow p_2 + p_2 \rightarrow p_1}{\sum_{i=1}^{N} p_1 \rightarrow p_i + \sum_{i=1}^{N} p_2 \rightarrow p_i} \quad (1)$$

3. Results

The average phone accuracy of each recogniser over all age groups is displayed in Table 2. As expected, the DNN systems outperform the GMM ones, with DNN2 scoring slightly lower than DNN1. As one would expect, the results are poorer than published phone recognition accuracies for adult speech. In addition, the difficulty of recognising children's speech will have been compounded by the use of a "flat" phone-level grammar.

The results for each individual age group confirm the classification of the four systems and show the anticipated gradual improvement in phone recognition accuracy with increasing age, starting from 56.6% for the first group in DNN1 and reaching 67% for the last group in the same system (Figure 2).

Less intuitively, but in accordance with previous findings, the total errors which can be predicted from PEALA present the same trend as the ASR accuracy: building up as the speakers grow older (Figure 2). Indicatively, in the case of DNN1, the percentage of PEALA related errors rose from 12% for five-year-olds to 18% for fifteen-year-olds, suggesting a correlation with phone accuracy.

In summary, moving to a DNN-HMM-based phone recognition system has resulted in improved phone recognition accuracy, but the tendency for the proportion of PEALA-related phone substitutions to increase with phone recognition accuracy, and hence with age, persists. Clearly this is the opposite of what one would expect and suggests that these errors are not attributable to factors associated with language acquisition but instead are due to the inherent confusibility of these pairs of phones.

As pointed out in a previous paper, this hypothesis is consistent with the fact that attempts to modify pronunciation dictio-

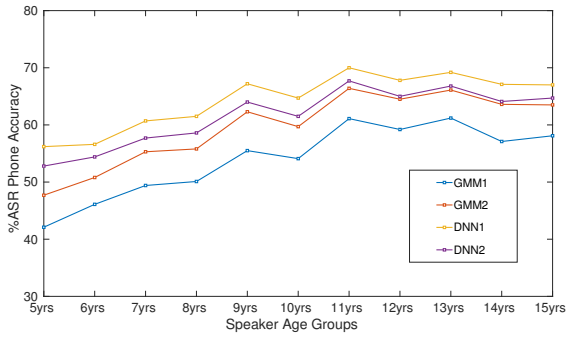

Figure 1: *Percentage phone accuracy as a function of age*

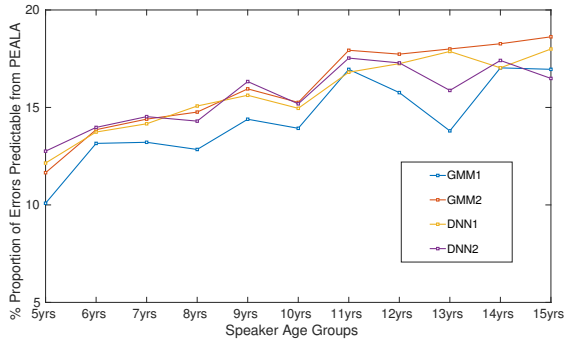

Figure 2: *Percentage of errors predictable from PEALAs as a function of age*

naries to accommodate mispronunciations in children's speech due to factors associated with language acquisition do not yield significant improvements.

In the absence of evidence for a significant effect of factors associated with language acquisition on phone recognition accuracy, we turned to an investigation of whether these phenomena are evident in the acoustic data.

Table 2: *Overall ASR Results per System*

	GMM1	**GMM2**	**DNN1**	**DNN2**
Mean Acc.	54.1%	59.8%	64.4%	61.6%
S. D.	9.6%	9.9%	9%	8.9%

4. Visualization

It was shown in [22] that low dimensional projections of BNFs can represent speech sounds in a topology that broadly reflects their phonetic properties, and that variations due to different initializations of the DNN may be compensated by suitable linear transformations. Motivated by this we hypothesized that a BNF induced image might offer an insight on how speech sounds evolve as a function of age.

In order to be able to visualise the BNFs, dimension reduction was attained by LDA resulting in sets of 2 dimensional vectors. For each phone, an age specific Gaussian ellipse was computed and plotted in different combinations of graphs. Due to the high level of overlap among the data, the phone cluster contours were chosen to correspond to 0.1 standard deviations allowing the phone clusters to be more easily distinguished.

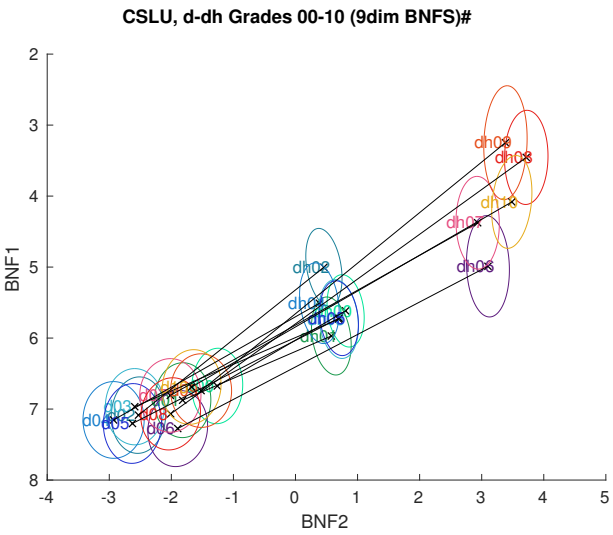

Figure 3: *d-dh. Gaussian ellipses are labelled according to the age group they represent. For example d00 corresponds to the youngest, 5-year-old, speaker group and d10 to the oldest, 15-year-old, speaker group.*

The Gaussian ellipses were colour-coded to represent the different age groups, with shades of green and blue for the younger speakers and red and yellow for the older ones.

Figure 3 shows such a plot for the confusable pair $/d/ - /dh/$ from table 1. For each age, a line connects the mean values of BNF features for $/d/$ and $/dh/$. There is a trend for the distance between the mean values of $/d/$ and $/dh/$ to increase with age. The observation that the $/dh/$ ellipses are closer to the $/d/$ ellipses for the youngest children is consistent with the hypothesis that they move progressively away from $/d/$ with increasing age. The figure shows a distinct cluster for $/d/$ in the bottom left-hand corner. The values for $/dh/$ form two age-dependent clusters, which separate the data in a non gradual manner in two groups; one below and one above age twelve. This distinct progression of age-dependent clusters could be a consequence of boys' voices breaking around that age, combined with the fact that $/dh/$ has been reported [8] to be one of the last consonants to be acquired and thus is more linguistically challenging. Indeed, this pattern of distinct age clusters was again observed in the data in the cases of $/v/$, $/r/$ and $/l/$ which are also among the lastly acquired phonemes. In the contrary, phonemes that appear first in the phonemic acquisition repertoire such as $/p/$, $/b/$ and $/d/$ [8] progress more gradually and in more compact clusters.

The distances between the mean values of substitution pairs featured in Table 1 were calculated for each age group and plotted as bar charts. These charts illustrate whether the phones in question come closer together or move away from each other as the speaker age increases. Following visual inspection these charts were divided approximately into 4 categories: (a) "increasing distance as a function of age", (b) "gently increasing distance", (c) "constant distance" and (d) "decreasing distance as a function of age". The confusable pairs falling into each category are indicated in Figure 4. If phone separation increases with age one would expect the majority of pairs to be in category (a) or (b), which is the case. The pair $/d/ - /dh/$, corresponding to Figure 3, falls into category (a). In general, plosives are featured in category (a) for the most part with a few

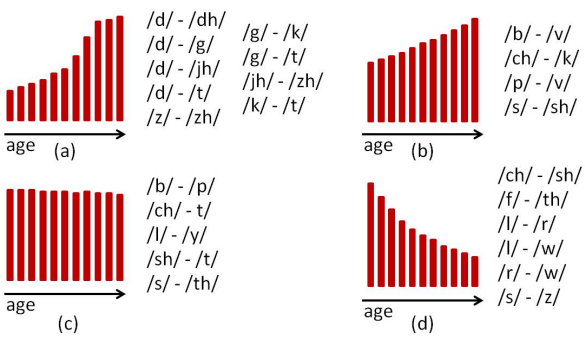

Figure 4: *Categorization of confusable pairs from Table 1 according to how the distances between pairs change with age. (a) "Increasing", (b) "gently increasing", (c) constant" and (d) decreasing". The bar charts are schematic and do not correspond to actual measurements*

appearances in categories (b) and (c), while liquids and glides are mostly featured in category (d) with one exception appearing in (c). Fricatives are spread throughout the four categories. This partition could potentially reveal some motif of linguistic development, however this would require further data analysis extending beyond the PEALA related pairs and needs to be revisited in a separate study.

Figure 5, corresponding to the pair $/b/ - /p/$, is an example from Category (c), where the separation of the phones is approximately constant and independent of age. Figure 5 suggests that there is considerable variability in the realizations of $/b/$ and $/p/$. It is also interesting to note the evolution of these realizations as a function of age. For example, the realizations of $/p/$ describe a semi-circle, converging for Grade 7 and above.

Figure 6, corresponding to $/f/ - /th/$ is an example from Category (4), where, contrary to expectation, separation *decreases* with increasing age. The clusters for $/f/$ and $/th/$ are diverse and overlapping compared with Figures 3 and 5. The trend for $/f/ - /th/$ to become more confusable with increasing age is supported by the results of the ASR experiment.

If this figure indicates a true phenomenon, that young teenagers' production of $/f/$ and $/th/$ becomes closer with age, then it seems very unlikely that this behaviour can be attributed to language acquisition and a different explanation needs to be found.

5. Conclusions

This work focuses on phonological effects associated with language acquisition in the context of automatic speech recognition. Four recognisers are developed on a relatively large children's speech corpus, offering a comparison across different training methods. The emerging phone confusion results replicate previous findings [5] and confirm the suggestion that as the system performance gets better, the proportion of phone substitutions predictable from PEALA gets higher. This is in contrast with how the proportion of PEALA evolve in the course of a speaker's development, therefore we conclude that these ASR confusions cannot be attributed to factors associated with language acquisition. We note that this conclusion is consistent with the observation that attempts to improve speech recognition performance for children by introducing mispronunciations due to language acquisition into the pronunciation dictionary do not deliver significant improvements in recognition accuracy.

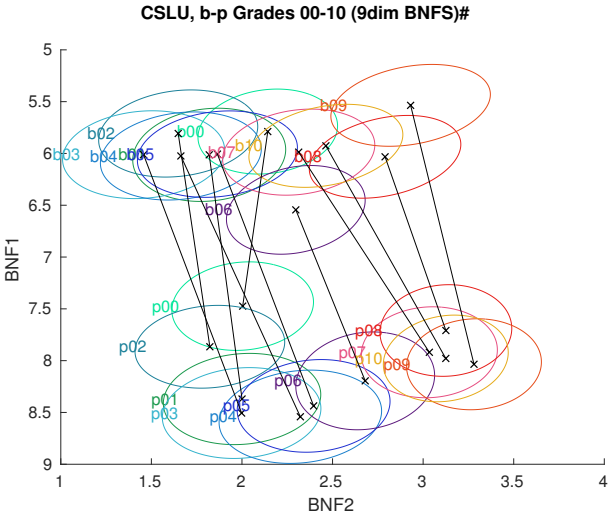

Figure 5: *b-p.*

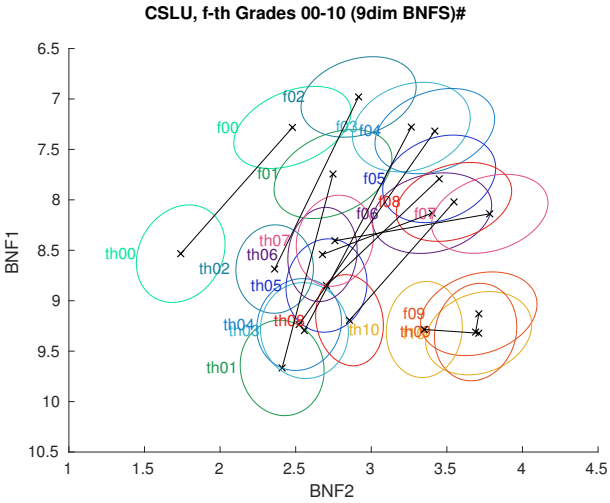

Figure 6: *f-th.*

In the absence of evidence for an effect of language acquisition factors on speech recognition performance, we investigated whether there is evidence of these factors in the acoustic data. Inspired by the findings in [22] we used low dimensional DNN bottleneck features to visualize changes in the acoustic realizations of phones as a function of age, focusing on 27 phone pairs whose confusability is potentially attributable to language acquisition factors. In the majority of cases we observed that, as expected, separation between the two classes increases with age. However, there are also some pairs for which separation remains relatively constant and, surprisingly, others such as $/f/ - /th/$ for which separation appears to decrease with age. We speculate that the latter phenomenon is due to factors other than language acquisition. Future work could extend the analysis of BNF representations beyond the PEALA related pairs.

6. References

[1] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[2] ——, "Developmental acoustic study of american english diphdongs," in *INTERSPEECH 2014*, 2014, pp. 1–16.

[3] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, "Analysing children's speech: An acoustic study of consonants and consonant-vowel transition," in *ICASSP 2006*, vol. 1, 2006, pp. 393–396.

[4] A. Holm, S. Crossbie, and B. Dodd, "Differentiating normal variability from inconsistency in childrens speech: normative data," *International Journal of Language and Communication Disorders*, vol. 42, no. 4, pp. 467–486, 2007.

[5] E. Fringi, J. Lehman, and M. Russell, "The role of phonological processes and acoustic confusability in phone errors in children's asr," in *WOCCI 2016 - Workshop on Child Computer Interaction*, 2016, pp. 10–15.

[6] D. Ingram, "Phonological rules in young children," *Journal of Child Language*, vol. 1, no. 1, pp. 49–64, 1974.

[7] ——, "Phonological development: Production." in *Language Acquisition: Studies in First Language Development*, P. Fletcher and M. Garman, Eds. Cambridge: Cambridge University Press, 1986.

[8] B. Dodd, A. Holm, Z. Hua, and S. Crossbie, "Phonological development: A normative study of british-english speaking children," *Clinical Linguistics and Phonetics*, vol. 17, no. 8, pp. 617–643, 2003.

[9] B. Smit, A., J. J. Hand, L. andFreilinger, E. Bernthal, J., and A. Bird, "The iowa articulation norms project and its nebraska replication," *Journal of Speech and Hearing Disorders*, vol. 55, pp. 779–798, 1990.

[10] P. Grunwell, "The development of phonology: A descriptive profile," *First Language*, vol. 2, no. 6, pp. 161–191, 1981.

[11] V. Hazan and S. Barrett, "The development of phonemic categorization in children aged 6+12," *Journal of Phonetics*, vol. 28, pp. 377–396, 2000.

[12] R. Romeo, V. Hazan, and M. Pettinato, "Developmental and gender-related trends of intra-talker variability in consonant production," *Acoustical Society of America*, vol. 134, no. 5, pp. 3781–3792, 1974.

[13] R. D. Kent, "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies," *Journal of Speech, Language, and Hearing Research*, vol. 19, pp. 421–447, 1976.

[14] B. Walsh and A. Smith, "Articulatory movements in adolescents," *Journal of Speech, Language, and Hearing Research*, vol. 45, pp. 1119–1133, 2002.

[15] A. Smith and H. Zelaznik, "Development of functional synergies for speech motor coordination in childhood and adolescence," *Developmental Psychobiology*, vol. 45, no. 1, pp. 22–33, 2004.

[16] S. McLeod and J. Arciuli, "School-aged children's production of /s/ and /r/ consonant clusters," *Folia Phoniatrica et Logopaedica*, vol. 61, pp. 336–341, 2009.

[17] W. Cohen and C. Anderson, "Identification of phonological processes in preschool children's sinlge-word productions," *International Journal of Language and Communication Disorder*, vol. 46, no. 4, pp. 481–488, 2011.

[18] E. Fringi, J. Lehman, and M. Russell, "Evidence of phonological processes in automatic recognition of children's speech," in *INTERSPEECH 2015*, 2015, pp. 1621–1624.

[19] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The ogi kids speech corpus and recognizers," in *ICSLP 2000 - Sixth International Conference on Spoken Language Processing*, vol. 4, 2000, pp. 258–261.

[20] D. Povey, A. Ghoshal, A. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 - Workshop on Automatic Speech Recognition and Understanding*, 2011.

[21] L. Bai, P. Jancovic, M. Russell, and P. Weber, "Analysis of a low-dimensional bottleneck neural network representation of speech for modelling speech dynamics," in *INTERSPEECH 2015*, 2015, pp. 583–587.

[22] P. Weber, L. Bai, M. Russell, P. Jancôviĉ, and S. Houghton, "Interpretation of low dimensional neural network bottleneck features in terms of human perception and production," in *Proc. Interspeech 2016*, 2016.